# 基于 SupervisedML 的方法流程与结果分析（大纲）

## 方法流程

1. 数据准备  
- 已知的耐药突变作为正样本（Positive）。  
- 未知/未报道的突变作为未标注样本（Unlabeled）。  
- 形成物种 × 突变的二进制矩阵（X\_dense）。

2. 模型选择与训练  
- 采用 Random Forest 作为监督学习的基分类器。  
- 通过 PU-learning 框架，将正样本与未标注样本结合建模。  
- 训练过程中随机遮蔽部分已知突变，用于评估模型的恢复能力。

3. 验证策略：Mask-then-Recover  
- 将部分已知突变遮蔽。  
- 使用模型预测候选突变。  
- 计算 Recall@K，评估在前 K 个预测中能否找回被遮蔽的突变。

4. 候选突变生成  
- Top-K 策略：每个物种取前 K 个预测概率最高的突变。  
- 阈值策略：取预测概率超过预设阈值的突变。  
- 生成候选突变清单，用于后续分析与实验验证。

## 结果分析

1. 模型评估  
- Recall@K 随 K 增加而上升，表明模型能较好恢复已知突变。  
- 在 K=20 时 Recall 达到约 0.85，之后趋于平台期，说明模型稳定。

2. 候选突变预测  
- Top-K 和阈值策略均能产生合理的候选集。  
- 部分候选突变在多个物种中出现，提示可能的保守耐药位点。  
- 预测结果为实验验证提供了优先级参考。

3. 生物学意义  
- 预测得到了一批文献未报道的潜在耐药突变。  
- 可作为后续体外实验和功能验证的候选目标。  
- 强调跨物种的一致性与差异性。

4. 方法优势与局限  
- 优势：在正负样本不均衡的情况下，PU-learning 能充分利用未标注数据。  
- 局限：预测结果仍需实验验证，数据偏倚可能影响模型的稳健性。